Motivational Interviewing
TRAINING · RESEARCH · IMPLEMENTATION · PRACTICE
MITRIP

Should We Trust our Judgments about the Proficiency of Motivational Interviewing Counselors?

A Glimpse at the Impact of Low Inter-rater Reliability

Chris Dunn, PhD,[1] Doyanne Darnell, PhD,[1] Sheng Kung Michael Yi, PhD,[2] Mark Steyvers, PhD,[2] Kristin Bumgardner, BS,[1] Sarah Peregrine Lord, PsyD,[1] Zac Imel, PhD,[3] David C. Atkins, PhD[1]

Abstract

Standardized rating systems are often used to evaluate the proficiency of Motivational Interviewing (MI) counselors. The published inter-rater reliability (degree of coder agreement) in many studies using these instruments has varied a great deal; some studies report MI proficiency scores that have only fair inter-rater reliability, and others report scores with excellent reliability. How much can we to trust the scores with fair versus excellent reliability? Using a Monte Carlo statistical simulation, we compared the impact of fair (0.50) versus excellent (0.90) reliability on the error rates of falsely judging a given counselor as MI proficient or not proficient. We found that improving the inter-rater reliability of any given score from 0.5 to 0.9 would cause a marked reduction in proficiency judgment errors, a reduction that in some MI evaluation situations would be critical. We discuss some practical tradeoffs inherent in various MI evaluation situations, and offer suggestions for applying findings from formal MI research to problems faced by real-world MI evaluators, to help them minimize the MI proficiency judgment errors bearing the greatest cost.

Keywords

motivational interviewing, inter-rater reliability, proficiency judgments, counselor proficiency, Motivational Interviewing Treatment Integrity

To assess counselor proficiency of the Motivational Interviewing (MI) intervention in a clinical trial (e.g., Moyers, Manuel, Wilson, Hendrickson, Talcott, & Durand, 2008), detect improvement in counselor proficiency in an MI training study (e.g., Baer, Wells, Rosengren, Hartzler, Beadnell, & Dunn, 2009), or admit MI counselors into the renowned Motivational Interviewing Network of Trainers (MINT) annual training, MI evaluators use one of several standardized MI scoring instruments such as the Motivational Interviewing Treatment Integrity scale (MITI; Moyers, Martin, Manual, Miller, & Ernst, 2009) or the Motivational Interviewing Skills Code (MISC; Miller, 2000). These measures require a trained person (coder) to listen to and score audio recordings of MI sessions according to a specific set of criteria designed to reflect counselor adherence to and competence in the delivery of MI. To be considered a reliable measure of MI counselor skill, two or more coders must score MI counseling sessions similarly, which is known as inter-rater reliability. Inter-rater reliability tells us how confident we can be that scores given by a coder for a specific MI criteria are accurate, with more confidence at greater degrees of inter-rater reliability. This study explores the impact of the degree of inter-rater reliability on making accurate versus flawed judgments about a counselor's proficiency in MI.

To illustrate the practical implications of inter-rater reliability, we present the example of judging the proficiency of the MITI scores of candidates seeking admission to the MINT. In the past, as part of their application to the MINT, candidates have each submitted an audio recording of an MI session that was scored using the MITI. The MITI is the standardized MI rating system most often used in scientific MI literature (Moyers, Martin, Manuel, Hendrickson, & Miller, 2005). The MITI measures counselor adherence and competence in MI with 5 global scores that characterize MI relevant aspects of the entire session and 6 counselor behavior counts, or tallies of counselor behavior over the course of the session that are relevant to MI. Global scores and behavior counts may then be combined into summary scores for which expert opinions are established for proficiency cut-offs (Moyers, Martin, Manual, Miller, & Ernst, 2009). Only those applicants' scores at or above the cut-offs would be judged as proficient scores, and these judgments were used by the MINT admissions committee as a key factor in accepting or rejecting applicants.

So how reliable is the MITI? Several psychometric studies of the MITI, each of which used different coding teams, have reported widely varying inter-rater reliability estimates for each MITI criteria (Bennett, Roberts, Vaughan, Gibbin, & Rouse 2007; Brueck, Frick, Loessl, Kriston, & Schondelmaier, 2009; Forsberg, Kallmen, Hermansson, Berman, & Helgason, 2007; Moyers, Martin, Manuel, Hendrickson, & Miller, 2005; Pierson, Hayes, Gifford, Roger, Padilla, Bissett, Kohlenberg, Rhode, & Fisher, 2007), both within and across studies. Variability in the reliability of

[1]Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, Seattle, WA, USA,
[2] Department of Cognitive Sciences, University of California, Irvine,CA, USA
[3] Department of Counseling Psychology, University of Utah, Salt Lake City, UT, USA

This research was supported by the National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health under award number R01AA018673. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The authors report no conflicts of interest.

Correspondence concerning this article should be to: Chris Dunn, Ph.D., Harborview Medical Center, Box 359911, 325 Ninth Avenue; Seattle, WA 98104. Email: cdunn@uw.edu

ISSN 2160-584X (online) Vol. 1 No. 3 (2014) DOI 10.5195/mitrip.2014.43

measuring MI with the MITI translates into variability in how confident we can be in the accuracy of MITI scores and, therefore, the accuracy of judgments made about counselors' proficiency in MI. Before presenting our methodology for exploring the impact of inter-rater reliability on judgments about counselor proficiency in MI, a brief explanation of inter-rater reliability and how it is estimated for the MITI is warranted.

To be useful, measurement tools must be reliable, or yield the same results on repeated assessments. In the case of coded data, this means two or more coders scoring the same counselor sessions must be able to come to the same conclusions about what scores or codes to assign for each counselor utterance that is coded. This is known as inter-rater reliability. To estimate inter-rater reliability for the MITI, two or more coders must first be sufficiently trained to score counseling sessions using the MITI, which is an iterative process of observing and resolving discrepancies in how coders are interpreting and using the scoring rules. Coders then each score the same sample of counseling sessions. Ostensibly, coders are recognizing characteristics of the counselor's MI skill. In the language of classical test theory (Lord & Novick, 1968), a characteristic of MI skill would be represented by a "true" MITI score or a "true" code, which can be thought of as the underlying best score that would be obtained from a consensus of MITI scoring experts if they were to conduct a hypothetical debate among themselves about the best score to assign to a particular case. In reality, we cannot actually observe the "true" scores. The best we can do in these types of measurement is to compare the MITI scores given by multiple coders for the same MI counseling sessions to make our best guess possible about how well the scores by the coders represent the "true" scores.

One method of quantifying how well the scores given by the coders represent the "true" MITI scores is by computing the Intraclass Correlation Coefficient (ICC), a statistic that describes how much of the total variation in MITI scores is due to differences between the coders (Shrout & Fleiss, 1979). The ICC is the preferred method of calculating inter-rater reliability for quantitative, continuous measures like scores on the MITI. ICCs are calculated for each MITI criteria and range from 0 to 1. Relatively high ICCs closer to 1 mean that coders scored the MITI in a similar way (i.e., coder scores correlated with each other relatively well). ICCs closer to 0 mean poor reliability and that there is considerable error due to how coders perceive and code the counselor MI skill. Cicchetti's (1994) guidelines for assessing the clinical significance of inter-rater reliability estimates, including for ICCs are "poor" < 0.40, "fair" = 0.40-0.59, "good" = 0.60-0.74, and "excellent" ≥ 0.75. If the inter-rater reliability estimate for a given MITI criteria is 0.40, this would mean that coders were quite inconsistent in how they scored the criteria and we would not have much confidence in the accuracy of those scores. Going back to the MINT application process as an example, if a given MITI score had poor reliability, then the chances of incorrectly judging that score as proficient/not proficient is high. If inter-rater reliability is low, this is generally an indication that either the measure is defective or the coders should be re-trained.

METHODS

If one knows the M, SD, the reliability of the observed MITI scores, and the cut-off to mark proficiency, it is possible to simulate example data to explore how reliability impacts judgments about counselor MI proficiency. The advantage of simulating data is that we directly control the properties of the data and thus can specifically examine the influence of reliability on the accuracy of judgments about proficiency. Specifically, we are able to create "true" scores, something we can never know in real data, and observed scores based on specified levels of reliability, and then compare these true and observed scores. We examined two different reliability levels, 0.50 and 0.90, representing the typical range of MITI reliability estimates found in the MI literature. For each set of values, a dataset with 10,000 counselor scores on one MITI summary score was simulated. Although there would never be a dataset with 10,000 counselors, this size guarantees that any error due to the simulation itself, so-called Monte Carlo error, is effectively zero.

To simplify the simulations, we focused on only one of five summary scores put forth by Moyers and colleagues (2009), the Reflection-to-Question Ratio (R/Q). The R/Q score combines 4 separate MITI scores including the number of 1) simple reflections, 2) complex reflections, 3) closed questions, and 4) open questions. It reflects the proportion the counselor's utterances in the session that were statements to demonstrate empathy for (or understanding of) the client versus questions posed to the client. Theoretically, making more reflections than questions in a session is consistent with good MI practice. The expert-derived cut-off score indicating proficiency in R/Q is 1. Both observed and true R/Q scores were simulated using the observed R/Q scores from 441 MI sessions (M = 1.71, SD = 1.60) in a large study of MI for drug abuse in primary care (Krupski, Joesch, Dunn, Donovan, Bumgardner, Peregrine Lord, Ries, & Roy-Byrne, 2012). Observed scores were simulated with error by systematically varying the level of reliability (i.e., fair vs. excellent). Because the observed R/Q scores were notably skewed, the simulations were based on log-transformed scores and then back-transformed to their

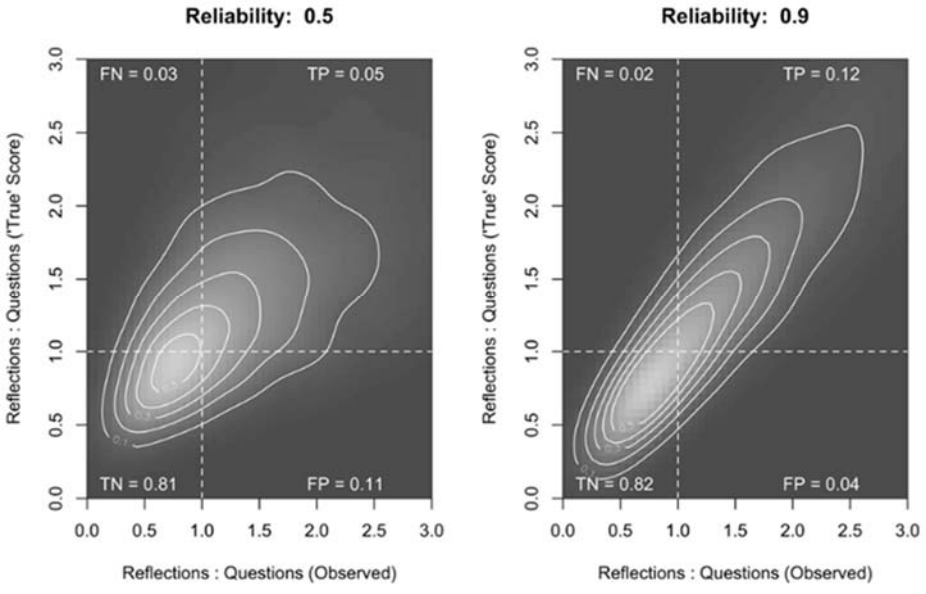

Figure 1

Simulated true scores versus observed scores for R/Q ratio for reliability of 0.50 and 0.90

1. Lighter shading indicates higher density of scores. The correlation of observed and true scores varies with the square root of the inter-rater reliability. Observed and "true" MI Spirit scores shown on x- and y-axes, respectively. Dotted lines represent proficiency cut-offs. False Positive (FP or falsely judging a score as proficient) and False Negative (FN or falsely judging a score as not proficient) judgments occur in lower right and upper left quadrants, respectively.
2. Each contour line indicates the proportion of data that is outside that ring (e.g., 0.10 is the widest circle, meaning that 10% of the total data is outside that ring).
3. The sample mean (1.71) and S.D. (1.60) used for this simulation was from Roy-Byrne et al. (2014).

ISSN 2160-584X (online) Vol. 1 No. 3 (2014) DOI 10.5195/mitrip.2014.43

original scale. All analyses including simulations were done using R v3.0.2 (R Development Core Team, 2013).

RESULTS

Figure 1 is an image plot with contour lines that displays simulated observed and true R/Q scores for 10,000 counselors with reliabilities of 0.50 and 0.90 and M = 1.71 and SD = 1.60. The horizontal axis represents observed R/Q scores, and the vertical axis represents "true" R/Q scores. The dashed vertical and horizontal lines show the cut-off score for judging a given counselor's score as proficiency on R/Q (score = 1) based on observed scores (vertical) and true scores (horizontal). All scores in the lower right quadrant are notated as False Positives (FP) because they fall above the cut-off for observed scores but are actually below the cut-off for true scores. FPs represent the situation in which MI counselors would be judged proficient on R/Q but truly are not. All scores falling in the upper left quadrant are notated as False Negatives (FN) because they fall below the cut-off for observed scores but actually fall above the cut-off for true scores. FNs represent the situation in which MI counselors would be judged as not proficient in R/Q but truly are. The shading represents the density of data points, with lighter grey indicating greater density. In addition, the contour lines convey more specific information about density, where each contour line encircles a specific proportion of the data. The number on each line describes what proportion of the data lies outside that contour line. For example, the largest contour line contains 90% of the data within its bounds (and consequently, 10% outside).

Figure 1 shows that the dispersion of scores is more narrow when reliability is at 0.90 than at 0.50, which means the observed and true scores are more similar when reliability is higher. Figure 1 shows that when reliability improves from 0.50 to 0.90, the FP rate (falsely judging a score as proficient) drops from 11% to 4%, and the FN rate (falsely judging a score as not proficient) drops from 3% to 2%. Consider the impact of reliability in the hypothetical situation in which 100 counselors applying for admission to the MINT had a mean R/Q of 1.71 with SD = 1.60 (as used in this simulation). If the ICC for their R/Q scores had been only 0.50, 11% or 11 of the 100 applicants would be falsely judged as proficient in R/Q. If the ICC had instead been 0.90, this number would have dropped to 4 applicants. In the case of FNs, 3 versus 2 applicants would be falsely judged as not proficient for reliabilities of 0.50 and 0.90, respectively.

DISCUSSION

Our aim was to examine and illustrate the impact of reliability on judgments of counselor proficiency in MI using Monte Carlo simulations, given that different coding teams in the MITI literature achieve widely varying inter-rater reliability results using the same instrument. The variation in inter-rater reliability should not be surprising given the different languages, MITI trainers, coders, counselors, clients, behavior change topics, or even sound quality of audio recordings across these studies, all of which add error to measurement. Nevertheless, rigorous MI measurement is crucial to enable us to trust our judgments of counselor MI proficiency and ensure the integrity of MI as it is widely implemented.

For the *M* and *SD* values used in this study, we see that the difference in "error savings" between fair and excellent reliability is greater for FPs (falsely judging a counselor as proficient) than in FNs (falsely judging a counselor as not proficient). Specifically, when reliability is 0.90 versus 0.50, the FP rate drops from 11% to 4%, and the FN rate drops from 3% to 2%. As depicted in Figure 1, it is clear that this is in part due to the fact that the *M* of R/Q of the sample used to create our simulated data sets was 1.70, well above the cut-off for deeming a counselor proficient in MI, making any FNs rare. In a sample with a lower *M* of R/Q, FNs would become more of an issue.

As to the question of what level of reliability is adequate for making judgments about counselor proficiency in MI, Cicchetti's 1994 paper ascribing meaning to levels of reliability as either poor, fair, good, or excellent is almost always cited in the MI literature. However, Cicchetti's earlier paper (1981) urging investigators to decide on what degree of reliability, and therefore what degree of confidence in the scores, would be sufficient for their specific purposes is seldom cited. Our simulation demonstrates that poor reliability may be more or less of a problem, depending on the MI evaluator's goals. As can be seen from Figure 1, altering the cut-off score over which counselors are judged proficient in R/Q would alter the number of FPs and FNs. MI evaluators may choose to alter the cut-off score to lower or raise the chance of committing each type of error (FPs/FNs), depending on which type of error is deemed to be the "less of two evils" (Charter, 2001). In the MINT application example, the organization can decide whether a FP or FN error bear the greatest cost to candidates, the organization, and the welfare of future learners trained by these candidates. If a FN were considered more harmful than a FP, MINT evaluators could lower proficiency cut-off scores so as to commit relatively more FPs and fewer FNs. Of course, how well certain cut-off scores actually translate into counselor proficiency in MI is also an important consideration, since greater MI skill should then translate into better client behavior change outcomes. Unfortunately, there is no available empirical guidance regarding the most appropriate proficiency cut-off scores for improving patient outcomes.

A related consideration is that the closer to the mean a given score is, the more at-risk the MI evaluator is of making a FP or FN error, that is, the less one can trust one's proficiency judgment. For example, an R/Q score of 0.95 is very close to the beginning proficiency cut-off of 1.00) but would still be deemed not proficient. Judging a score that were say, 0.30 or 1.70, would be an easier call to make. When it is a close call, and the score's reliability is low, evaluators should put less weight on that score. For instance, the MINT organizers would have less certainty about judgments of MI proficiency for counselors close to the R/Q cutoff and may then decide to give less weight to these applicants' R/Q scores in the application process, instead attending to other aspects of their applications to decide whether to admit the counselor to MINT.

Judgments of counselor MI proficiency in real-world dissemination settings where coder training may be less intensive are probably even more prone to FN and FP errors. Training MITI coders is costly and time-consuming, and probably better in the context of well-funded research studies. However, even among well-funded studies we see considerable variability in the reliability of MITI scores. It is likely that reliability is even lower in real-world settings that are unlikely to be able to afford a similar degree of rigorous training. Additionally, within counselors, there is known to be wide variation in skill across multiple MI sessions (Forsberg, Forsbert, Lindqvist, & Helgason, 2010). Coding multiple work samples should improve the accuracy of judgments made about counselor proficiency in MI (Imel, Baldwin, Baer, Hartzler, Dunn, Rosengren, & Atkins, 2013). We recommend that supervisors using a tool like the MITI to score supervisee MI sessions use caution when considering any single score from any single session as indicative of counselor skill in MI.

A limitation of this study is that we examined only a single MITI score (R/Q), when in fact the MITI has five summary scores that are commonly used to evaluate counselor proficiency in MI. The complexity of the simulation methodology used in this study increases beyond feasibility when even one more score is added. Although the same tradeoffs between the costs of high reliability versus judgment error rates (FPs/FNs) would pertain to any of the five MITI proficiency scores, MI evaluators still have little to go on when they must somehow combine all 5 summary scores to arrive at a judgment about counselor proficiency they can trust. Ideally, in the future, a weighted mean proficiency formula might be

ISSN 2160-584X (online) Vol. 1 No. 3 (2014) DOI 10.5195/mitrip.2014.43

developed to allow evaluators to attach weights to each proficiency score, given these other factors.

We have tried to help MI evaluators to more systematically consider the consequences of using data reflecting counselor MI proficiency with questionable reliability. As long as reliability itself is relatively unreliable, there can be no substitute for careful human judgment in evaluating counselor proficiency in MI. We have used the MITI to illustrate the importance of reliability, however, the implications of our simulations also apply to other standardized MI rating systems in use (Miller, 2000; Ball, Martino, Corvino, Morganstern, & Carroll, 2002; Martino, Ball, Nich, Frankforter, & Carroll, 2008).

REFERENCES

Baer, J. S., Wells, E. A., Rosengren, D. B., Hartzler, B., Beadnell, B., & Dunn, C. (2009). Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors. *Journal of Substance Abuse Treatment, 37*, 191-202.

Ball, S.A., Martino, S., Corvino, J., Morganstern, J., Carroll, K.M., 2002. Independent tape rater guide. Unpublished psychotherapy tape rating manual.

Benett, G.A., Roberts, H.A., Vaughan, T.E., Gibbin, J.A., & Rouse, L. (2007) Evaluating a method of assessing competence in motivational interviewing: A study using simulated patients in the United Kingdom. *Addictive Behaviors*, 32, 69-79.

Brueck, R.K., Frick, K., Loessl, B., Kriston, L., & Schondelmaier, S. (2009). Psychometric properties of the German version of the Motivational Interviewing Treatment Integrity Code. *Journal of Substance Abuse Treatment*, 36, 44-48.

Charter, R.A., & Feldt, L.S. (2001). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical and Experimental Neuropsychology*, 23, 530-537.

Cicchetti, D.V., & Sparrow, S.A. (1981). Developing criteria for establishing inter-rater reliability of specific items: Applications to assessment of adaptive behavior. *Journal of Mental Deficiency.* 86,127-137.

Cicchetti, D.V. (1994). Guidelines, criteria and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284-290.

Forsberg, L., Forsbert, L.G., Lindqvist, H., & Helgason, A.R. (2010). Clinician acquisition and retention of Motivational Interviewing skills: A two-and-a-half-year exploratory study. *Substance Abuse Treatment, Prevention, and Policy,* 5:8, 1-14.

Forsberg, L., Kallmen, H., Hermansson, U., Berman, A.H., & Helgason, A.R. (2007). Coding counselor behavior in motivational sessions: Inter-rater reliability for the Swedish Motivational Interviewing Treatment Integrity Code (MITI). *Cognitive Behavior Therapy*, 36, 162-169.

Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Imel, Z.E., Baldwin, S., Baer, J., Hartzler, B., Dunn, C., Rosengren, D., & Atkins, D. (2013). Evaluating therapist competence in motivational interviewing by comparing performance with standardized and real patients. 82(3), 472-481.

Lord, F.M., & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Welsley Publishing Company.

Lord, S.P., Atkins, D., Imel, Z., Dunn, C. (2013). Advancing methods for data collection and reliability of the Motivational Interviewing Skills Code. pii: S0740-5472(14)00188-3. doi: 10.1016/j.jsat.2014.08.005. [Epub ahead of print].

Madson, M.B., Loignon, A.C., & Lane, C. (2009). Training in motivational interviewing: A systematic review. *Journal of Substance Abuse Treatment* 36, 67-73.

Martino, S., Ball, S.A., Nich, C., Frankforter, T.F., & Carroll, K.M. (2008). Community program counselor adherence and competence in motivational enhancement therapy. *Drug and Alcohol Dependence, 96,* 37-48.

Miller, W.R. (2000). Motivational Interviewing Skill Code (MISC): Coder's manual. Unpublished manual: University of New Mexico.

Moyers, T.B., Manuel, J.K., Wilson, P.G., Hendrickson, S.M.L., Talcott, W., & Durand, P. (2008). A randomized trial investigating training in motivational interviewing for behavioral health providers. *Behavioural and Cognitive Psychotherapy*, 36, 149-162.

Moyers, T.B., Martin, T., Manual, J.K., Miller, W.R., & Ernst, D. (2009). Revised global scales: Motivational Interviewing Treatment Integrity 3.1 (MITI 3.1). Available from http://casaa.unm.edu/download/MITI3_1.pdf .

Moyers, T.B., Martin, T., Manuel, J.K., Hendrickson, S.M.L., & Miller, W.R. (2005). Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment*, 28, 19-26.

Pierson, H.M., Hayes, S.C., Gifford, E.V., Roger, N., Padilla, M., Bissett, R., Kohlenberg, B., Rhode, R., & Fisher, G. (2007). An examination of the Motivational Interviewing Treatment Integrity Code. *Journal of Substance Abuse, 32, 11-17.*

Roy-Byrne, P., Bumgardner, K., Krupski, A., Dunn, C., Ries, R., Donovan, D., West, I., Maynard, C., Atkins, D.C., Graves, M., Joesch, J., & Zarkin, G.A. (2014). Brief intervention for problem drug use in safety-net primary care settings: A randomized clinical trial. *Journal of the American Medical Association, 312(5):492-501.*

Shrout, P. and Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Swets, J.A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, 47, 522-532.

ISSN 2160-584X (online) Vol. 1 No. 3 (2014) DOI 10.5195/mitrip.2014.43